SUPPLEMENT TO "ESTIMATION OF THE ANNUAL SURVIVAL RATE OF A STATIONARY POPULATION" *

BU-107-M D. S. Robson June 23, 1959

A Test of the Model

The estimation problem discussed in BU-106-M is frequently complicated in practice by an inability on the part of the experimenter to obtain a random sample from a natural population. The sample is obtained, in the case of animal populations, by capture of the animals, and the method of capture is ordinarily selective for certain age groups. Fish samples, for example, are usually obtained by a netting operation which may be ineffective at capturing the very small fish of the younger age groups, resulting in a sample deficit of young fish relative to older fish. Fishery biologists conventionally circumvent this difficulty by using in their estimation procedure only the sample data from the older age groups which are fully vulnerable to the method of capture employed; the youngest age group which is fully vulnerable to capture was labeled the 0-age group in previous sections (BU-106-M). There is, however, no way of knowing exactly where this dividing line between incomplete and complete vulnerability falls, so the data of the youngest age group used in estimation of survival rate is always suspect. We propose now to construct a procedure for testing the validity of the geometric model, with particular emphasis on the validity of the data of the 0-class.

For our test criterion, then, we seek a statistic which is a function of the number, say N_0, of sample members falling in the 0-class. A well-known statistic satisfying this condition is the so-called "Jackson Estimate" of the survival rate s,

$$s' = \frac{N_1+N_2+\cdots+N_k}{N_0+N_1+\cdots+N_{k-1}} = \frac{n-N_0}{n-N_k}$$

where N_i is the number of sample members falling in the i'th age group and k is the oldest age group represented in the sample. This estimator is widely used in practice though it is not well defined and, unfortunately, can take on values greater than 1. A modified form of this estimator,

$$s_0' = \frac{N_1+N_2+\cdots+N_k}{N_0+N_1+\cdots+N_k} = \frac{n-N_0}{n}$$

*The numbering of the equations is continued from BU-106-M

however, is well defined and is, in fact, unbiased. The numerical value of s_0' will usually not differ greatly from s' since the sample number of elements N_k in the oldest age group will usually be small, but s_0' has the satisfying property $0 \le s_0' \le 1$. Moreover, for fixed n, s_0' is a monotonic function of the chance variable N_0 and so appears very suitable as a test statistic for the present purposes.

The distribution of the statistic s_0' clearly depends upon the unknown parameter s; however, the sufficiency of the statistic $T=\Sigma X_i$ implies that the conditional distribution of s_0' for a given T is independent of s. We compute this conditional distribution by first noting that the joint conditional distribution of $X_1,\cdots,X_n$ for a given T is

$$\Pr(X_1=x_1,\cdots,X_n=x_n \mid T=t)=\begin{cases}\binom{n+t-1}{t}^{-1} & \text{for } (x_1,\cdots,x_n \mid \Sigma x_i=t)\\ 0 & \text{otherwise}\end{cases}$$

that is, every ordered set $(x_1,\cdots,x_n \mid \Sigma x_i=t)$ of non-negative integers is equally likely. The number of such ordered sets containing exactly N_0 zeros, N_1 ones, N_2 twos, $\cdots$, and N_t t's is simply a multinomial coefficient, so

$$(3)\qquad \Pr(N_0=n_0,\cdots,N_t=n_t \mid \Sigma N_i=n,\Sigma i n_i=t)=\frac{n!}{n_0!\cdots n_t!}\binom{n+t-1}{t}^{-1}$$

Verification that (3) is a probability distribution is obtained by equating the coefficients of s^t on the left and right hand sides of

$$(1+s+s^2+s^3+\cdots)^n=(1-s)^{-n}$$

The conditional distribution of N_0 is then

$$(4)\qquad \Pr\left\{N_0=n_0 \mid \Sigma N_i=n,\Sigma i N_i=t\right\}=\binom{n}{n_0}\sum_{\substack{n_1,\cdots,n_t \\ \Sigma_1^t n_i=n-n_0 \\ \Sigma i n_i=t}}\frac{(n-n_0)!}{n_1!\cdots n_t!}\binom{n+t-1}{t}^{-1}$$

and, as seen by equating coefficients of s^t on the left and right hand sides of

$$(s+s^2+s^3+\cdots)^{n-n_0}=s^{n-n_0}(1-s)^{-(n-n_0)}$$

the distribution (4) is the hypergeometric distribution

$$\Pr(N_0=n_0 \mid \Sigma N_i=n, \Sigma i N_i=t) = \frac{\binom{n}{n-n_0}\binom{t-1}{t-n+n_0}}{\binom{n+t-1}{t}}$$

(5)

$$= \Pr(n-N_0=n-n_0 \mid \Sigma N_i=n, \Sigma i N_i=t$$

The earlier assertion that the modified "Jackson Estimate" $s_0'=(n-N_0)/n$ is unbiased is now readily verified, since

$$E\left(\frac{n-N_0}{n}\middle| T\right) = \frac{T}{n+T-1} = \hat{s}_T$$

and, as previously noted, $E\hat{s}=s$. The variance of s_0' is also easily computed in the form

$$\operatorname{var}(s_0') = E\left\{\operatorname{var}(s_0' \mid T)\right\} + \operatorname{var}(\hat{s})$$

$$= E\left\{\frac{T(T-1)(n-1)}{n(n+T-1)^2(n+T-2)}\right\} + \operatorname{var}(\hat{s})$$

The second component of this variance, $\operatorname{var}(\hat{s}_T)$, as indicated earlier is approximately

$$\operatorname{var}(\hat{s}_T) \doteq \frac{s(1-s)^2}{n}$$

and the first component may be approximated by

$$E\left\{\operatorname{var}(s_0' \mid T)\right\} \doteq E\left\{\frac{T(T-1)(n-1)}{n(n+T-1)(n+T-2)(n+T-3)}\right\} = \frac{s^2(1-s)}{n} .$$

Added together, these two approximations give

$$\operatorname{var}(s_0') \doteq \frac{s^2(1-s)}{n} + \frac{s(1-s)^2}{n} = \frac{s(1-s)}{n}$$

Oddly enough, the errors in these two approximations cancel each other when added together, giving a final answer which is exactly correct. The unconditional distribution of the statistic $n-N_0$ is, in fact, a binomial distribution with parameter s; compounding the distribution (5) with the distribution (2) of

T we find

$$\Pr(n-N_0=n-n_0)=\sum_{t=0}^{\infty}\Pr(N-n_0=n-n_0|T=t)\Pr(T=t)$$

$$=\binom{n}{n-n_0}\sum_{t=0}^{\infty}\binom{t-1}{t-n+n_0}s^t(1-s)^n$$

$$=\binom{n}{n-n_0}s^{n-n_0}(1-s)^{n_0}$$

so that

$$\operatorname{var}\left(\frac{n-N_0}{n}\right)=\frac{s(1-s)}{n}$$

An exact test of the validity of the model then consists of comparing the observed statistic s_0' with the critical values, say $s_{.025}$ and $s_{.975}$, computed from the hypergeometric distribution (4). Because of the close approximation of the binomial to the hypergeometric distribution, however, such test procedures are ordinarily replaced in practice by the binomial test. This approximate test procedure would then consist of entering a table of binomial confidence intervals (or Clopper-Pearson charts) with "sample size"=T and "number of successes"=$T-n+N_0$; if the resulting confidence interval covers $\hat{s}$ then the validity of the model is accepted, and otherwise it is rejected. For large samples, of course, this procedure may be replaced by the normal test; for fixed T the statistic

$$\frac{\frac{n-N_0}{n}-\frac{T}{n+T-1}}{\sqrt{\frac{T(T-1)(n-1)}{n(n+T-1)^2(n+T-1)}}}=\frac{s_0'-\hat{s}}{\sqrt{\operatorname{var}(s_0'\,T)}}$$

has mean 0 and variance 1 and is asymptotically distributed as a standard normal deviate. The approach of the hypergeometric to the normal distribution is rapid, and with the sample sizes commonly employed in this type of experiment, say $n>100$, little is to be gained by using the more exact binomial or hypergeometric test procedures; furthermore, the availability of extensive tables for the normal distribution permits experimenters to make one-tailed tests of the hypothesis of a deficit in the 0-class.

Other Estimators of s

If the assumptions of the geometric model (1) are satisfied then for most practical purposes the estimator $\hat{s}_T$ would be considered optimum. Among all unbiased estimators it has the smallest variance or, more generally, for any convex loss function the average loss associated with $\hat{s}_T$ is less than that of any other unbiased estimator. The requirement that an estimator be unbiased, however, is not really a defensible one, and when this requirement is lifted the number of worthy (admissible) estimators becomes unlimited. Furthermore, within the class of unbiased estimators in which only $\hat{s}_T$ is admissible there may be some which are actually superior to $\hat{s}_T$ in the practical sense that their desirable properties are less dependent upon the rigid set of assumptions which lead to the sufficiency and completeness of T. For these and purely academic reasons we proceed now to examine some other possible estimators of s.

The statistic s_0' has already been considered because of its relation to the estimator s' now in use. Another modification of s' which has been used in practice is the ratio

$$\frac{N_2+\cdots+N_k}{N_1+\cdots+N_{k-1}} = \frac{n-N_0-N_1}{n-N_0-N_k}$$

in which the data from the zero-class has been omitted. Again, this ratio is undefined with positive probability under model (1), but an analogous well defined unbiased estimator is

$$s_1' = \begin{cases} \dfrac{n-N_0-N_1}{n-N_0} & \text{if } T>1 \\ \dfrac{1}{n} & \text{if } T=1 \\ 0 & \text{if } T=0 \end{cases}$$

The conditional distribution of N_1 given both N_0 and T, $T>1$, is the hypergeometric

$$\Pr(N_1=n_1 \mid N_0=n_0, T=t) = \frac{\binom{n-n_0}{n-n_0-n_1}\binom{t-(n-n_0)-1}{t-2n+2n_0+n_1}}{\binom{t-1}{t-(n-n_0)}}$$

and if T=1 then, of course, $N_1=1$ with probability 1. For $T>1$ the conditional expectation of s_1' is

$$E\left(\frac{n-N_0-N_1}{n-N_0} \,\middle|\, N_0=n_0, T=t\right) = \frac{t-n+n_0}{t-1}$$

and

$$E\left(\frac{T-n+N_0}{T-1} \,\middle|\, T=t\right) = \frac{t}{n+t-1} \quad .$$

Consequently, for all T, $E(s_1'|T)=\hat{s}_T$. This method of construction of unbiased estimators may be extended to give s_2', s_3', etc., but their definitions become increasingly cumbersome.

Another estimator used in practice is the average of the ratios $N_1/N_0, N_2/N_1, \cdots, N_k/N_{k-1}$ where k is the oldest age group in the sample or some preassigned maximum age for which data is used. This estimator, like s', is not well defined for all possible outcomes of the experiment, but a slight modification such as adding 1 to each denominator will remedy this. Estimators of this form are used even in situations where the survival rate is known to vary with age; i.e., where the proportion s_0 of individuals surviving their first year is known to be different from the proportion s_1 surviving from their first to second birthday, and so on. The ratio N_i/N_{i-1} is, in fact, the maximum likelihood estimator of s_i under the model

$$f(x)= \frac{s_1 s_2 \cdots s_x}{1+s_1+s_1 s_2+\cdots+s_1 \cdots s_k} \tag{6}$$

so that

$$\bar{s}' = \frac{1}{k}\left(\frac{N_1}{N_0} + \cdots + \frac{N_k}{N_{k-1}}\right)$$

is the maximum likelihood estimator of the average annual survival rate $\frac{1}{k}(s_1+\cdots+s_k)$. The vector $(N_0, N_1, \cdots, N_k)$ rather than $T=\Sigma i N_i$ is the sufficient statistic in this case, and the distribution of this vector is the multinomial

$$\Pr\left\{N_0=n_0, \cdots, N_k=n_k\right\} = \frac{n!}{n_0! \cdots n_k!} \; \frac{s_1^{n_1}(s_1 s_2)^{n_2} \cdots (s_1 \cdots s_k)^{n_k}}{(1+s_1+s_1 s_2+\cdots+s_1 \cdots s_k)^n}$$

We can construct an estimator analogous to $\bar{s}$ which is unbiased under model (1) using the fact that for $t>0$

$$E\left(\frac{N_{i+1}}{N_i+1}\middle| T=t\right)= \frac{t}{n+t-1} - \frac{\sum_{\upsilon=0} (-1)^{\upsilon}\binom{n}{\upsilon}\binom{n+t-2-\upsilon(i+1)}{t-1-\upsilon i}}{\binom{n+t-1}{t}}$$

$$= \frac{t}{n+t-1}\left[1-\Pr(N_i=0\middle| T=t-1)\right]$$

So for $i=0,1,\cdots,T-1$ the estimator

$$\bar{s}_i = \frac{N_{i+1}}{(N_i+1)\left[1-\Pr(N_i=0|T=t-1)\right]}$$

if $T>0$ and $\bar{s}_0=0$ if $T=0$, though somewhat ridiculous in form, is unbiased. The average of these $\bar{s}_i$,

$$\bar{s} = \begin{cases} \frac{1}{T}(\bar{s}_0+\cdots+\bar{s}_{T-1}) & \text{if } T > 0 \\ 0 & \text{if } T = 0 \end{cases}$$

is then an unbiased analogue of the estimator $\bar{s}'$.

A noteworthy difference between the two estimators $\bar{s}'$ and $\bar{s}$ is that the former represents an average of some fixed number k of ratios while the latter is an average of T ratios, where T is a chance variable. If the number of age groups to be used in estimation is fixed in advance, as it presumably is in practice when the estimator $\bar{s}'$ is employed, then the model is effectively truncated on the right, and in order to examine the properties of $\bar{s}'$ or its analogues for the case of constant mortality rates it becomes necessary to truncate model (1) at some preassigned age x=k. The truncated model (1), which is then strictly analogous to model (6), will be considered in the next section; first, however, another case should be mentioned where the number k represents the oldest age group occurring in the sample.

Where the estimator $\bar{s}'$ has been used in the literature it is not always explicitly stated whether the number k was fixed in advance or determined by the data in some way. The distinction is of both theoretical and practical importance since the method of determining k determines the bias in the estimation

procedure. If k is determined by the sample data then, of course, k is a chance variable having a probability distribution which depends upon the exact rule used by the experimenter in arriving at k, and if this rule is not explicitly stated then the bias in the estimation procedure is unknown. When k represents the oldest age group occurring in the sample then under model (1) the distribution of $K=\max(X_1,\cdots,X_n)$ is

$$\Pr(K=k)=\Pr(N_i=0 \text{ for all } i>k)-\Pr(N_i=0 \text{ for all } i\geq k)$$
$$=(1-s^{k+1})^n-(1-s^k)^n$$

and the ratio $N_i+1/(N_i+1)$ has the conditional expectation

$$E\left(\frac{N_{i+1}}{N_i+1}\middle| K=k>0\right)=s\left[1-\frac{(1-s^{k+1}-s^i(1-s))^n-(1-s^k-s^i(1-s))^n}{(1-s^{k+1})^n-(1-s^k)^n}\right]$$

The average of these ratios is therefore an estimate of

$$(7)\qquad E\left(\frac{1}{K}\sum_{i=0}^{K-1}\frac{N_{i+1}}{N_i+1}\right)=s\left[1-\sum_{k=1}^{\infty}\frac{1}{k}\sum_{i=0}^{k-1}\left\{(1-s^{k+1}-s^i(1-s))^n-(1-s^k-s^i(1-s))^n\right\}\right]$$

if we adopt the convention that the statistic takes on the value 0 when K=0. In this form, the bias of the estimator may be compared with the bias of the corresponding, fixed k, estimator to be described in the next section on the truncated geometric model.

The Truncated Model

If the experimenter decides in advance that he will not use data on fish of age greater than k in his estimation procedure then the geometric model (1) is effectively truncated to

$$f(x)=\frac{s^x(1-s)}{1-s^{k+1}}\ ,\quad x=0,1,\cdots,k\ .$$

The joint distribution of n independent observations (ages) from this truncated distribution is

$$P_k(X_1=x_1,\cdots,X_n=x_n)=\frac{s^{\sum_1^n x_i}(1-s)^n}{(1-s^{k+1})^n}$$

showing, again, that $T=\sum_1^n X_i$ is a sufficient statistic. The distribution of T

obtained by equating coefficients of like powers of z in the identity

$$\sum_{t=0}^{nk} z^t P_k(t) \equiv \frac{[1-(zs)^{k+1}]^n}{(1-zs)^n} \quad \frac{(1-s)^n}{(1-s^{k+1})^n}$$

is given by

$$P_k(t) = \sum_{v=0}^{[\frac{t}{k+1}]} (-1)^v \binom{n}{v} \binom{n+t-1-v(k+1)}{n-1} s^t \left(\frac{1-s}{1-s^{k+1}}\right)^n$$

where $[\frac{t}{k+1}]$ is the integer part of $\frac{t}{k+1}$. Completeness of T follows from the fact that if $Eh(T)\equiv 0$ then the polynomial in s

$$\sum_{t=0}^{nk} \left\{ h(t) \sum_{v=0}^{[\frac{t}{k+1}]} (-1)^v \binom{n}{v} \binom{n+t-1-v(k+1)}{n-1} \right\} s^t$$

is identically 0, implying that h(t) is 0 for all t.

Completeness of T is not especially helpful in the truncated case, however, because the parameter s is no longer estimable; that is, no unbiased estimator of s exists. If there were some function of the observations, say $S(x_1,\cdots,x_n)$, such that

$$ES(X_1,\cdots,X_n)=s$$

then the conditional expectation, say h(t),

$$E\left\{ S(X_1,\cdots,X_n) \middle| \sum_1^n X_i = t \right\} = h(t)$$

would also be an unbiased estimator of s. This is impossible, however, for then

$$\sum_{t=0}^{nk} \left\{ h(t) \sum_{v=0}^{[\frac{t}{k+1}]} (-1)^v \binom{n}{v} \binom{n+t-1-v(k+1)}{n-1} \right\} s^t$$

$$\equiv_s \sum_{t=0}^{nk} \left\{ \sum_{v=0}^{[\frac{t}{k+1}]} (-1)^v \binom{n}{v} \binom{n+t-1-v(k+1)}{n-1} \right\} s^{t+1}$$

and this identity cannot hold because the coefficient of s^{nk+1} is 0 on the left hand side but non-zero on the right hand side.

Theoretically, a test of the model based on the zero-class frequency is still available for the truncated case since the distribution of N_0, given the sufficient statistic T, does not depend upon s. This conditional distribution assumes an awkward form, however, and has not been tabulated; it is presented here only to display its form. First, we obtain the conditional distribution of $N_0,\cdots,N_k$ as

$$P_k(N_0=n_0,\cdots,N_k=n_k \Big| \sum_{i=0}^{k} iN_i=t) = \frac{n!}{n_0!\cdots n_k!} \Big/ \sum_{\upsilon=0}^{[\frac{t}{k+1}]} (-1)^{\upsilon}\binom{n}{\upsilon}\binom{n+t-1-\upsilon(k+1)}{n-1}$$

verified by equating coefficients of s^t on the left and right hand sides of the identity

$$(1+s+s^2+\cdots+s^k)^n \equiv (1-s^{k+1})^n(1-s)^{-n}$$

Next, summing over $n_1,\cdots,n_k$, we get

$$(8) \qquad P_k(N_0=n_0|t)= \frac{\binom{n}{n_0} \sum_{\upsilon=0}^{[\frac{t-n+n_0}{k}]} (-1)^{\upsilon}\binom{n-n_0}{\upsilon}\binom{t-1-\upsilon k}{n-n_0-1}}{\sum_{\upsilon=0}^{[\frac{t}{k+1}]} (-1)^{\upsilon}\binom{n}{\upsilon}\binom{n+t-1-\upsilon(1+1)}{n-1}}$$

as may be verified from the identity

$$(s+s^2+\cdots+s^k)^{n-n_0} \equiv s^{n-n_0}(1-s^k)^{n-n_0}(1-s)^{-(n-n_0)}$$

The distribution (8), though not unmanageable for very small samples, appears to be too cumbersome for use with samples of practical size. If $k \geq t$ then (8) reduces to the hypergeometric distribution (5).

The unconditional distribution of the zero-class frequency is, again, a binomial distribution

$$P_k(n-N_0=r)=\binom{n}{r}\left[\frac{s(1-s^k)}{1-s^{k+1}}\right]^r\left[\frac{1-s}{1-s^{k+1}}\right]^{n-r}$$

so the statistic $(n-N_0)/n$, which was unbiased in the unrestricted model, now has expectation

$$E\left(\frac{N_1+\cdots+N_k}{N_0+\cdots+N_k}\right)=s\left[\frac{1-s^k}{1-s^{k+1}}\right] \quad .$$

The average of the ratios $N_{i+1}/(N_i+1)$ now has the expectation

$$E\left(\frac{1}{k}\sum_{i=0}^{k-1}\frac{N_{i+1}}{N_i+1}\right)=s\left[1-\frac{1}{k}\sum_{i=0}^{k-1}\frac{(1-s^{k+1}-s^i(1-s))^n}{(1-s^{k+1})^n}\right]$$

as compared to (7).

The most efficient estimator for the truncated model is the maximum likelihood estimator, obtained as the iterative solution to the likelihood equation

$$\frac{T}{n}=\frac{s}{1-s}-(k+1)\frac{s^{k+1}}{1-s^{k+1}}$$